Proceedings of the Fifth International Provenance and Annotation Workshop (IPAW), Cologne, Germany, June 9-13, 2014.

Challenges in Modeling Geospatial Provenance

Daniel Garijo[1], Yolanda Gil[2], and Andreas Harth[3]

[1] Ontology Engineering Group, Universidad Politécnica de Madrid
Boadilla del Monte, Madrid, Spain
[2] Information Sciences Institute, University of Southern California,
4676 Admiralty Way, Marina del Rey, CA 90292, USA
[3] Institute AIFB, Karlsruhe Institute of Technology
Englerstr. 11, 76128 Karlsruhe, Germany
dgarijo@fi.upm.es, gil@isi.edu, harth@kit.edu

Abstract. The surge in availability of geospatial data sources, the increased use of crowdsourced maps and the advent of geospatial mashups have brought us to an era where geospatial information is delivered to users after integration from diverse sources. Understanding the provenance of geospatial data is crucial for assessing the quality of the data and addressing whether to trust the information or not. In this paper we describe user requirements for modeling geospatial provenance.

Keywords: Provenance, geospatial data, metadata, data integration.

1 Introduction

The Open Geospatial Consortium and the World Wide Web Consortium are working jointly towards standards for linking and integrating geospatial data [Archer 2014]. As geospatial data is often used in decision making (e.g., navigation), the accuracy of integrated data is important. Assessing the correctness of information requires tracking the origins of the data. Geospatial data presents several challenges, which leads us to explore the following issues concerning provenance:

This paper presents a study on user requirements for geospatial provenance, based on discussions with users, researchers, and practitioners at several meetings and workshops on geospatial data. We are using these user requirements in our work to drive representations of geospatial provenance using W3C PROV recommendation [Moreau et al 2013].

2 User Requirements for Geospatial Provenance

We assume that the user is presented with a map that integrates data from several data sources. We assume the data sources to have published their content in the web of data, but the discussion applies also to traditional integration scenarios. Each dataset contains objects such as geospatial features (e.g., regions) with attributes (e.g.,

population) and links to other objects, such as geospatial geometries (e.g., the polygon of a region at a certain resolution). The goal of the integration process is to create an integrated map where all the attributes have one single value.

We make the following assumptions:

- **Each dataset contains objects with unique identifiers**. If two different versions datasets cover the same object, we assume the objects are mapped to each other to be able to derive a unique identifier.
- **The objects across datasets have been mapped**. The mapping step may be part of the integration process, but in our discussion we assume the mappings have been done. Note that the mappings across entities might contain errors.

Table 1. Questions from users about geospatial provenance.

PROVENANCE OF DATASETS:	**PROVENANCE OF SETS OF DATASETS:**
Q1: Where does the information in this map come from?	Q7: What maps were generated after a given date?
Q2: Who created the map?	Q8: Which maps were generated by a given organization/person?
Q3: How was the map created?	Q9: Which maps were generated with a given version of a source dataset?
Q4: What is the most recent version of this map?	Q10: Which maps were generated with a given version of the integration algorithm?
Q5: Why was the map updated?	
Q6: How was the map updated?	
PROVENANCE OF OBJECTS:	**PROVENANCE OF SETS OF OBJECTS:**
Q11: What original data source did this object come from?	Q16: What other objects in the map (or selected region) come from the same data source as a given selected object?
Q12: Who created the object?	Q17: What objects were taken from data from a given organization?
Q13: How was this object created?	Q18: What objects were taken from a specific original data source?
Q14: When was this object created?	Q19: What objects were taken from a type of data source (e.g., a crowdsourced data source)?
Q15: How was this object included in the original data source?	Q20: What objects were generated with an older version of the algorithm?
PROVENANCE OF PROPERTIES:	**PROVENANCE OF SETS OF PROPERTIES:**
Q21: What original data source did this property come from?	Q26: What properties of the selected objects come from the same data source as the selected property of that object?
Q22: Who created the property?	Q27: What properties of the selected objects
Q23: How was this property created?	Q28: What properties of a selected objects were taken from a specific original data source?
Q24: When was this property created?	Q29: What properties of a selected objects were taken from a type of data source (e.g., a crowdsourced data source)?
Q25: How was this property included in the original data source?	Q30: What properties were generated with an older version of the algorithm?
	Q31: What properties from other objects come from the same data source as a given selected property of an object?
OTHER PROVENANCE QUESTIONS:	
Q32: How did the selected information come about in each of the input data sources?	
Q33: How did a given set of manual corrections help improve later versions of the map?	
Q34: What is new in this new version of the map?	
Q35: What objects were integrated with confidence > 0.8?	
Q36: Why is the object I am looking for not appearing?	
Q37: Which datasets were used for generating a selected area?	
Q38: Can I see some highlights of important things about this map, e.g., where is the information more uncertain, where is the information really recent, where has the information changed the most, etc?	

- **The datasets share the same data model and vocabulary**. We assume the source datasets use the same object types and properties. For example, if one dataset used "latitude" and the other "lat", those properties have already been mapped by an upstream process. That mapping is a separate integration process that could be described using similar mechanisms to what is discussed in this document.

A user looking at a map might naïvely believe that all information is equal in quality. This is not a good assumption, as the quality of the information shown depends highly on the quality of the sources, the quality of the algorithm, and many other factors. Our goal is to help users understand the information they are seeing in a map so they can determine whether to trust it. We define trust as a judgment that a user makes based on the context of the information they see [Artz and Gil 2007]. A crucial part of this context is provenance, which aims to capture who/what/when/how/why the information was generated. Therefore, provenance information is crucial to provide context for users to make trust decisions.

We have collected provenance-related questions that would help a user assess their trust on a map and the information it contains. These questions were raised by users and based on discussions with users, researchers, and practitioners at several workshops on geospatial data.

Table 1 summarizes user requirements in terms of questions that would require geospatial provenance.

We are using these questions in our current work as requirements to drive representations of geospatial provenance using the W3C PROV standard.

Acknowledgements. This research was supported in part by the US National Science Foundation with award # IIS-1117281, by the Spanish Science and Innovation Ministry (MICINN) with an FPU grant (Formación de Profesorado Universitario), by ONR Global under grant no N62909-13-1-N024, and by the EU's FP7 programme via the PlanetData Network of Excellence (grant agreement no 257641).

References

Archer, P., et al. Joint W3C/OGC Workshop on Linking Geospatial Data, March 2014. http://www.w3.org/2014/03/lgd/

Artz, Donovan and Yolanda Gil. "A survey of trust in computer science and the Semantic Web." Journal of Web Semantics, 5(2):58-71. 2007.

Moreau, L.; Missier, P.; Belhajjame, K.; B'Far, R.; Cheney, J.; Coppens, S.; Cresswell, S.; Gil, Y.; Groth, P.; Klyne, G.; Lebo, T.; McCusker, J.; Miles, S.; Myers, J.; Sahoo, S.; and Tilmes, C. "PROV-DM: The PROV Data Model." World Wide Web (W3C) Recommendation, April 2013.